DOCUMENT RESUME

ED 093 974 TM 003 829

AUTHOR	Thrash, Susan K.; Porter, Andrew C.
TITLE	Invalidity of a Current Method for Estimating Reliability.
PUB DATE	[74]
NOTE	12p.; Paper presented at National Council on Measurement in Education Annual Meeting (Chicago, Illinois, April, 1974)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Comparative Analysis; Correlation; Measurement Techniques; *Sampling; *Testing Problems; *Test Reliability

ABSTRACT

The purpose of this paper is to prove that one currently recommended method of obtaining the reliability of an instrument defined on a population of aggregate units is invalid. This method randomly splits the aggregate into two halves, correlates the two half unit scores by a Pearson product moment correlation coefficient, and corrects the correlation coefficient using the Spearman-Brown prophecy formula. Our approach was to compare this procedure to the standard method of forming random split halves of items on the test. In addition the reliability of an instrument was obtained by both methods. It was found that the currently recommended method is an underestimate of the reliability of a test defined on an aggregate. (Author)

INVALIDITY OF A CURRENT METHOD FOR ESTIMATING RELIABILITY

Susan K. Thrash

Michigan State University
and
Department of Civil Service
State of Michigan

Andrew C. Porter

National Institute of Education

Paper presented at the 1974 NCME meetings

In educational research and evaluation the unit of analysis is frequently some aggregate of smaller units. A popular example is the use of classrooms, where an observation on a classroom is defined by some function of the observations on the students in the classroom. The purpose of the present paper is to consider the problem of estimating the reliability of a test attendant to the use of aggregate units. More specifically we prove that a currently recommended method of estimating the reliability of a test defined on a population of aggregate units is invalid. Our discussion is limited to the situation where individuals are measured by a uni-dimensional test and observations on aggregate units are defined by the mean of the observations of individuals comprising the aggregate units.

The paper proceeds by first considering the relationship between the reliability of a test for a population of aggregate units and for the population of individuals used to form those aggregate units. Next, we define the method of estimating reliability that is shown to be invalid. The analytic demonstration of invalidity is supplemented by a numerical example.

The reliability of a test for aggregate units

It is well known that the reliability of an instrument can vary across populations for which the instrument may be used. Even when the set of individuals is held constant the choice of unit of analysis represents a further definition of the population. It follows that for a given set of children, the reliability of a test for the population of children might well differ from the reliability of the same test for the population of classrooms in which the children experience their schooling. Similarly the reliability for this population of classrooms might differ from the reliability for the population of schools in which the classrooms are nested.

Shaycoft (1963) has investigated the relationship between the reliability of a test for a population of individuals and its reliability for a population of aggregate units formed by those individuals. She pointed out that the two reliabilities will be equal if the aggregate units are formed randomly. The reliability of a test for aggregate units will be greater than for individuals when the variance of the aggregate unit means is greater than what would be expected by random grouping. Although this is typically the case in education she goes on to say that the reverse will be true when the variance of the aggregate means is less than would be expected from random grouping. The size of the difference between the two reliabilities is a function of

1) the degree of departure from randomness,
2) the number of individuals in each aggregate,
3) the size of the reliability defined on individuals.

<u>The invalid method of estimating reliability</u>

The method to be considered for estimating the reliability of an instrument for a population of aggregate units can be described using schools as an example. First, randomly split each school into two halves and obtain a score on the instrument for each random half. Then, calculate the correlation between the two half unit scores by a Pearson product moment correlation coefficient. The reliability defined on schools is obtained by correcting the correlation coefficient using the Spearman-Brown prophecy formula.

Our first exposure to the above described method of estimating reliabilities was during the second author's participation in a consultant panel conference on the evaluation of the Follow Through Program. At that conference the method was suggested for estimating reliabilities of pretests where school was the unit of analysis. The reliabilities were needed for subsequent corrections to be made in analyses of covariance. Later we discovered that the procedure had been used, except for the part involving the Spearman-Brown correction, by Dyer, Linn, and Patton (1969) as a method for estimating the reliability of test defined on a population of school systems. O'Connor (1972) used Dyer, et al reliabilities in an example, but first corrected them using the Spearman-Brown formula to obtain estimates of the parallel forms reliabilities based on the full school systems. Since the procedure for estimating the reliability of a test defined on a population of aggregate units has enjoyed some popularity, it is of interest to investigate the properties of the procedure.

Analytic Demonstration

Our general approach was to compare the procedure for estimating reliability under investigation to the standard method of forming random split halves of items on the test, where a school's score on a split half of the test is the mean score for the students in the school. Where the two procedures are not in agreement the former is considered in error.

Starting with the split units procedure, the correlation between half unit scores on the full test is by definition

$$r_{X_1'X_2'} = \frac{\sum^{N} x_1'x_2'}{N\sigma_{X_1'}\sigma_{X_2'}}, \qquad (1)^*$$

where x_1' and x_2' are deviation half unit scores on the full test for the two sets of halves, $\sigma_{X_1'}$ and $\sigma_{X_2'}$ are the two standard deviations, and N is the number of units. Assuming the two standard deviations to be equal (which is the long run expectation) and that true scores and errors of measurement are independent for half units,

$$r_{X_1'X_2'} = \frac{\sum^{N} t_1't_2'}{N\sigma^2_{X_1'}} \qquad (2)$$

where t_1' and t_2' are true deviation half unit scores on the full test.

Further the correlation between the true half unit scores on the full test is by definition

$$r_{T_1'T_2'} = \frac{\sum^{N} t_1't_2'}{N\sigma_{T_1'}\sigma_{T_2'}},$$

*σ is not a parameter.

which simplifies to

$$r_{T_1'T_2'} = \frac{\sum^{N} t_1' t_2'}{N\sigma_{T_1'}\sigma_{T_2'}} \quad , \tag{3}$$

given the assumption that $\sigma_{T_1'} = \sigma_{T_2'}$. By way of equation (3), equation (2) becomes

$$r_{X_1'X_2'} = r_{T_1'T_2'} \frac{\sigma^2_{T_1'}}{\sigma^2_{X_1'}} \quad . \tag{4}$$

Now $\sigma^2_{T_1'}$ and $\sigma^2_{X_1'}$ need to be defined in terms of half test for full unit statistics. First, consider $\sigma^2_{X_1'}$. Letting X_1 and X_2 denote half-test scores for full units and assuming that the variances of the two half test scores on full units are equal, $\sigma^2_{X_1} = \sigma^2_{X_2}$, it follows that the variance of the full test for the full units is

$$\sigma^2_X = 2\sigma^2_{X_1} + 2r_{X_1X_2}\sigma^2_{X_1} \quad , \tag{5}$$

where $r_{X_1X_2}$ denotes the correlation between half-test scores for full units. But, the variance of the full test for full units is also

$$\sigma^2_X = 1/4(2\sigma^2_{X_1'} + 2r_{X_1'X_2'}\sigma^2_{X_1'}) \quad , \tag{6}$$

since

$$X = \frac{X_1' + X_2'}{2} \quad . \text{*}$$

*This is not exactly true for units comprised of an odd number of subunits.

Using equations (5) and (6)

$$2\sigma^2_{X_1} + 2r_{X_1X_2}\sigma^2_{X_1} = 1/4(2\sigma^2_{X'_1} + 2r_{X'_1X'_2}\sigma^2_{X'_1}) \quad ,$$

from which it follows that

$$\sigma^2_{X'_1} = \frac{4\sigma^2_{X_1}(1 + r_{X_1X_2})}{(1 + r_{X'_1X'_2})} \quad . \tag{7}$$

A similar strategy can be used to define $\sigma^2_{T'_1}$ in terms of half test full unit statistics. Letting T_1 and T_2 denote true half-test scores for full units and assuming that the variance of the two sets of true half test scores are equal, $\sigma^2_{T_1} = \sigma^2_{T_2}$, it follows that the variance of the true scores for the full test is

$$\sigma^2_T = 2\sigma^2_{T_1} + 2r_{T_1T_2}\sigma^2_{T_1} \quad , \tag{8}$$

where $r_{T_1T_2}$ is the correlation between T_1 and T_2. By classical measurement theory $r_{T_1T_2}$ equals one so that equation (8) becomes

$$\sigma^2_T = 4\sigma^2_{T_1} \quad . \tag{9}$$

But the variance of true scores for the full test is also

$$\sigma^2_T = 1/4(2\sigma^2_{T'_1} + 2r_{T'_1T'_2}\sigma^2_{T'_1}) \quad , \tag{10}$$

where again the prime indicates that the statistics are for half units on the full test. Using equations (9) and (10)

$$4\sigma^2_{T_1} = 1/4(2\sigma^2_{T'_1} + 2r_{T'_1T'_2}\sigma^2_{T'_1}) \quad ,$$

from which it follows that

$$\sigma^2_{T'_1} = \frac{8\sigma^2_{T_1}}{(1 + r_{T'_1T'_2})} \quad . \tag{11}$$

Returning to equation (4) and using the definitions provided by equations (7) and (11)

$$r_{X'_1X'_2} = \frac{r_{T'_1T'_2}\left(\frac{8\sigma^2_{T_1}}{1 + r_{T'_1T'_2}}\right)}{\frac{4\sigma^2_{X_1}(1 + r_{X_1X_2})}{1 + r_{X'_1X'_2}}} \quad ,$$

which reduces to

$$r_{X'_1X'_2} = \frac{2r_{T'_1T'_2}\sigma^2_{T_1}(1 + r_{X'_1X'_2})}{(1 + r_{T'_1T'_2})\sigma^2_{X_1}(1 + r_{X_1X_2})} \quad .$$

Solving for $r_{X'_1X'_2}$,

$$r_{X'_1X'_2} = \frac{2r_{T'_1T'_2}\sigma^2_{T_1}}{(1 + r_{T'_1T'_2})\sigma^2_{X_1}(1 + r_{X_1X_2}) - 2r_{T'_1T'_2}\sigma^2_{T_1}} \quad . \tag{12}$$

Since $r_{X_1X_2}$ is the reliability of the half test for full units it follows that

$$r_{X_1X_2} = \frac{\sigma^2_{T_1}}{\sigma^2_{X_1}} \quad . \tag{13}$$

Substituting the definition provided by equation (13) into equation (12)

$$r_{X_1'X_2'} = \frac{2r_{T_1'T_2'}r_{X_1X_2}}{(1 + r_{T_1'T_2'})(1 + r_{X_1X_2}) - 2r_{T_1'T_2'}r_{X_1X_2}} ,$$

which reduces to

$$r_{X_1'X_2'} = \frac{2r_{X_1X_2}r_{T_1'T_2'}}{2 - (1 - r_{X_1X_2})(1 - r_{T_1'T_2'})} . \qquad (14)$$

From equation (14) it follows that the two procedures for estimating reliability yield identical results when the correlation between the true half unit scores on the full test, $r_{T_1'T_2'}$, equals one. Given random splits on the units, $r_{T_1'T_2'}$ will equal one only when the standard error of the difference between the true score means of each pair of half units is zero. The standard errors have expectations greater than zero for schools of finite size. Thus for practical situations the split unit procedure does not yield results identical to the split test procedure.

Since we know that the estimation procedure under investigation is not in agreement with the standard, it is of interest to describe the nature of their lack of agreement. Our approach was to consider relative error (RERR) where RERR is defined as $\frac{\text{true estimate} - \text{new estimate}}{\text{true estimate}}$.

Defining $r_{X_1X_2}$ as the true estimate and $r_{X_1'X_2'}$ as the new estimate, we obtain

$$\text{RERR} = \frac{r_{X_1X_2} - \left[\dfrac{2r_{X_1X_2}r_{T_1'T_2'}}{2 - (1 - r_{X_1X_2})(1 - r_{T_1'T_2'})}\right]}{r_{X_1X_2}} .$$

This reduces to

$$\frac{r_{X_1X_2}(1 + r_{X_1X_2} - r_{T_1'T_2'} - r_{X_1X_2}r_{T_1'T_2'})}{r_{X_1X_2}[2 - (1 - r_{X_1X_2})(1 - r_{T_1'T_2'})]}$$

or

$$\text{RERR} = \frac{(1 - r_{T_1'T_2'})(1 + r_{X_1X_2})}{2 - (1 - r_{X_1X_2})(1 - r_{T_1'T_2'})} \quad . \qquad (15)$$

Note that when $r_{T_1'T_2'} = 1.00$, RERR = 0 which agrees with our earlier finding. In order to find when RERR is a maximum, we took the derivative with respect to $r_{X_1X_2}$. The values of $r_{X_1X_2}$ that make the derivative zero give the points of $r_{X_1X_2}$ where RERR is maximized. We found that there are no maximums or minimums except at the endpoints. Since $r_{X_1X_2}$ is bounded by 0 and 1, we found for $r_{T_1'T_2'} \geq 0$ that $\frac{1 - r_{T_1'T_2'}}{1 + r_{T_1'T_2'}} \leq \text{RERR} \leq 1 - r_{T_1'T_2'}$ and for $r_{T_1'T_2'} \leq 0$ that $\frac{1 - r_{T_1'T_2'}}{1 + r_{T_1'T_2'}} \geq \text{RERR} \geq 1 - r_{T_1'T_2'}$. Also since RERR is always positive for all values of $r_{T_1'T_2'}$ and $r_{X_1X_2}$, it follows that $r_{X_1'X_2'} \leq r_{X_1X_2}$.

Since for all practical situations the correlation between half unit scores for the full test has been shown to be less than the correlation between half-test scores for the full units, their Spearman-Brown corrected counterparts must maintain the same inequality. The conclusion is that the split units method provides an underestimate of the reliability of a test defined on a population of aggregate units.

<u>Example</u>

In order to illustrate the inequality of the two procedures for estimating the reliability of a test for a population of aggregate units, we used data on children in 35 classrooms ranging in size from 6 to 17 children. The basic data consisted of children's responses to the thirteen items on Part A of the Reading Subtest of the MAT Primary Level II, Form F. The children were second graders tested in the spring of 1973.

A table of random numbers was used to split each class into two halves, then half class means on the full test were calculated. The mean and variance of the half class means for one set of half classes were 6.29 and 3.02 respectively, while the mean and variance of the other set of half classes were 6.40 and 2.72 respectively. The mean equality of the two variances supports the practical utility of the corresponding assumption of equal variances made in the previous analytic demonstration. The correlation between the two sets of half class means was .17. The Spearman-Brown correction yields the value .29.

A table of random numbers was also used to split the test into two halves, then full class means on the half tests were calculated. The mean and variance of the full class means for one half of the test were 2.62 and .38 respectively, while the mean and variance for the other half of the test were 3.70 and .56 respectively. Again the two variances were nearly equal which supported the corresponding assumption made previously. For longer tests or tests with an even number of items the assumption of equal half test variances is even more likely. The correlation between the two half tests was .82 which became .90 using the Spearman-Brown correction.

Thus for the example the discrepancy between the two procedures for estimating reliability was substantial and in the predicted direction.

A secondary interest was to use the data to provide an example of the difference between the reliability of a test for aggregate units and the same test for the individuals comprising those aggregate units. Using the same split of the test as previously, the correlation between the two halves for children was .41 which became .58 using the Spearman-Brown correction.

Conclusions

When the unit of analysis is some aggregate unit, the reliability of a test should be reported for the population of aggregate units rather than for the population of individuals which form those units. In theory the size of the reliabilites for the two populations of units can differ in either direction, but in educational research the reliability defined on the population of aggregate units will typically be the larger.

The procedure of estimating the reliability of a test for aggregate units by forming split units, systematically underestimates the reliability and so should not be used. One acceptable method for estimating the reliability of a test for aggregate units parallels the familiar split test method. Shaycoft (1963) has provided other estimation procedures that are a function of the reliability of the test for the population of individuals on which the aggregate units are defined.

The utility of our finding can be illustrated by an example. When an educational researcher is attempting to "tease out" causal relationships where random assignment has not been employed, he sometimes uses partial correlations or estimated true scores analysis of covariance (Porter, 1973). For the former, the correlations of the variable being controlled with the other variables should be corrected for attenuation (Kahneman, 1963). For the latter the reliability of the covariate can be used in estimated true scores analysis of covariance (Porter, 1974). When the unit of analysis represents some aggregate of smaller units, the reliabilities used for the corrections should be defined on the population of aggregate units. The method investigated here would provide reliability coefficients which are too small and thus cause the statistical analyses to over-correct for the control variable.

References

Dyer, H. S., Linn, R. L., and Patton, M. J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. AERJ, 1969(6), 591-605.

Kahneman, D. Control of spurious association and the reliability of the controlled variable. Psychological Bulletin, 1965, 64, 326-329.

O'Connor, Test theory and the measurement of change, RER, Winter 1972, 42,1.

Porter, A. C. Analysis strategies for some common evaluation paradigms. Paper presented at the meetings of the American Educational Research Association, 1973.

Shaycoft, M. F. The statistical characteristics of school means. In Flanagan, J. C., Dailey, J. T., Shaycroft, M. F., Orr, D. B., and Goldberg, I. Studies of the American high school. (Final report to the U.S. Office of Education, Cooperative Research Project No. 226), Washington, D.C.: Project TALENT Office, University of Pittsburgh, 1962.

Shaycoft, M. F. The use of school means as variables. Revised version of a paper presented at the annual meetings of the American Psychological Association, Philadelphia, September, 1963.